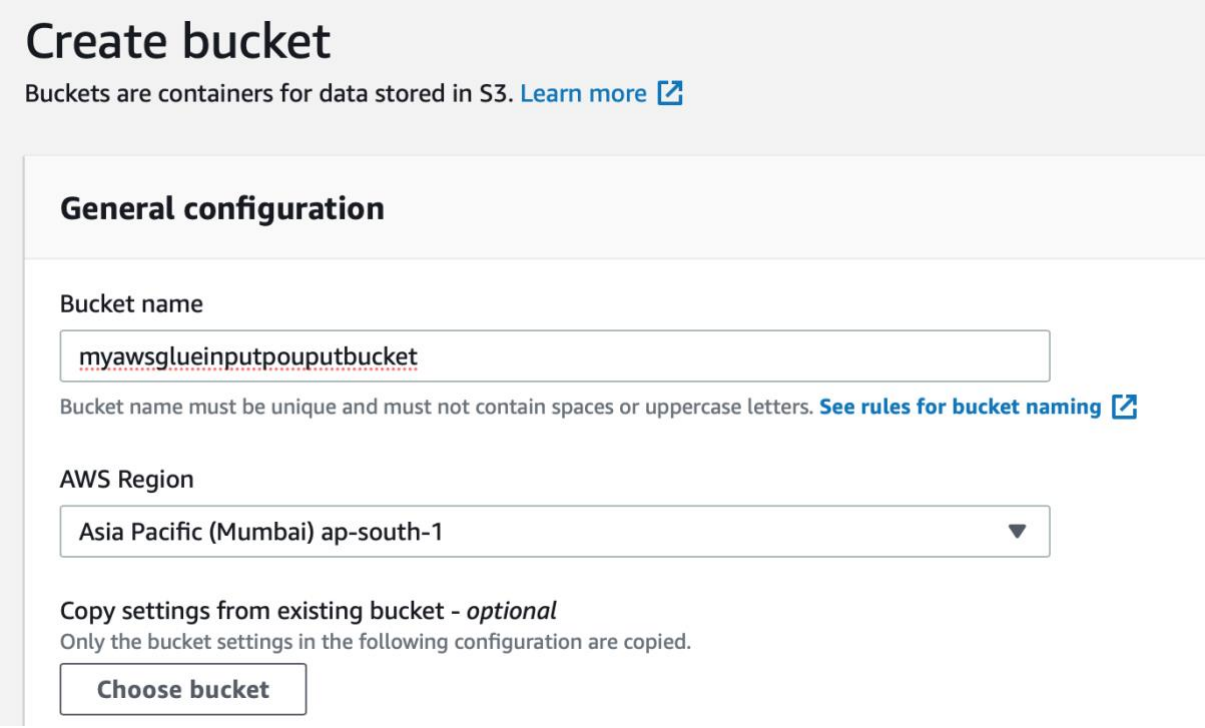
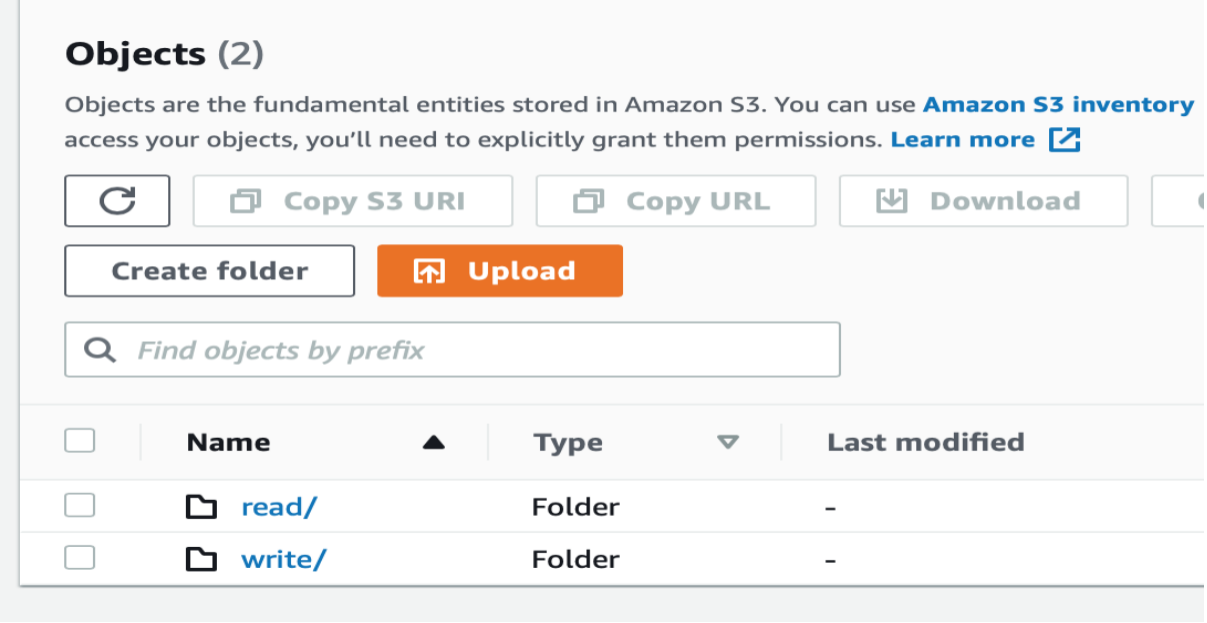
Exercise – Data Processing using AWS Glue

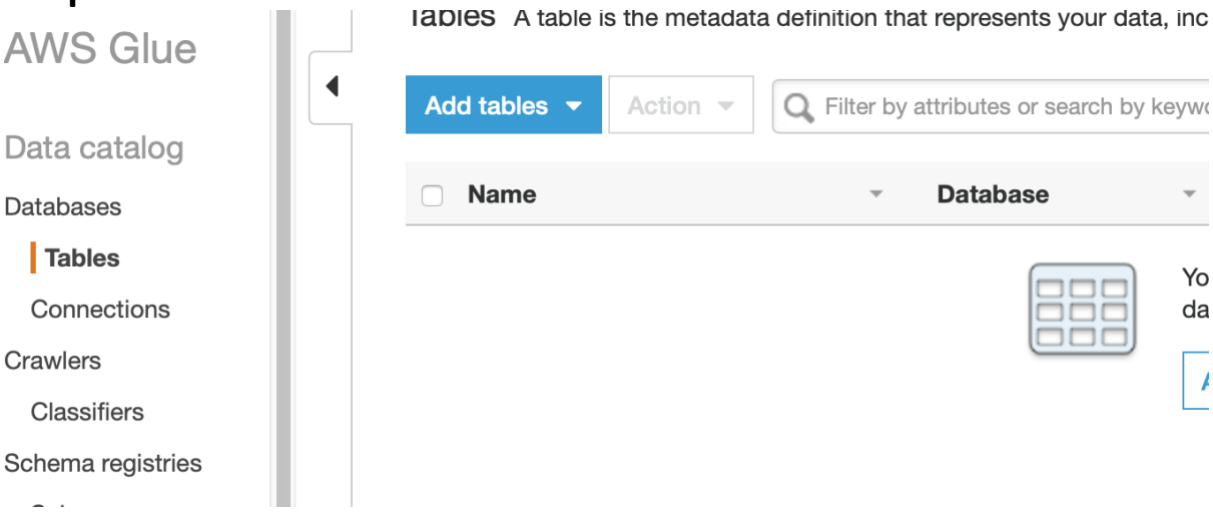
Step 1: Create a new bucket so that AWS Glue can read from and write to the bucket in to it's own folders.



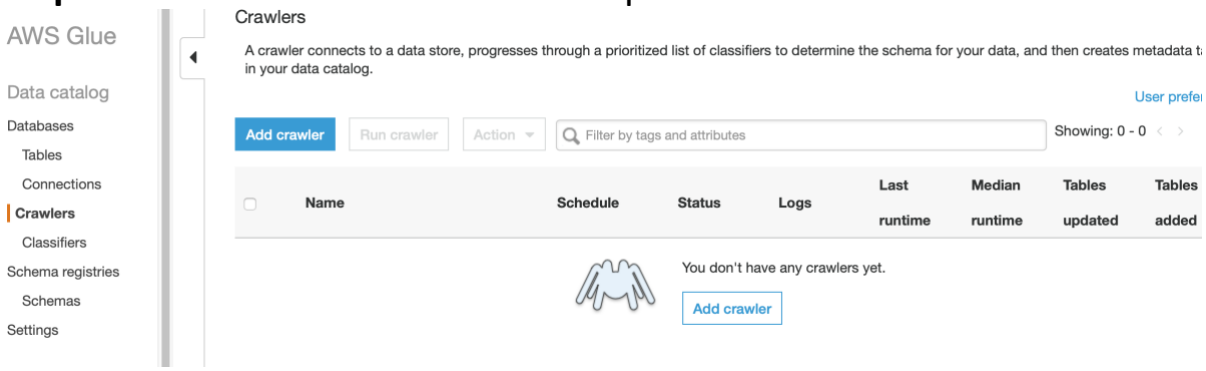
Step 2: Create a couple of folders one for reading and one for writing to/from Glue. Upload the movielens csv file to the read directory in S3.



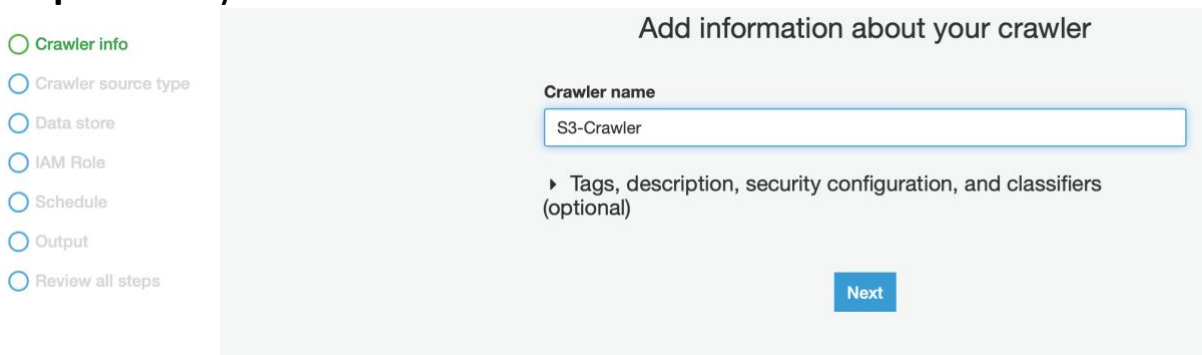
Step 3: Go to AWS Glue in Amazon console



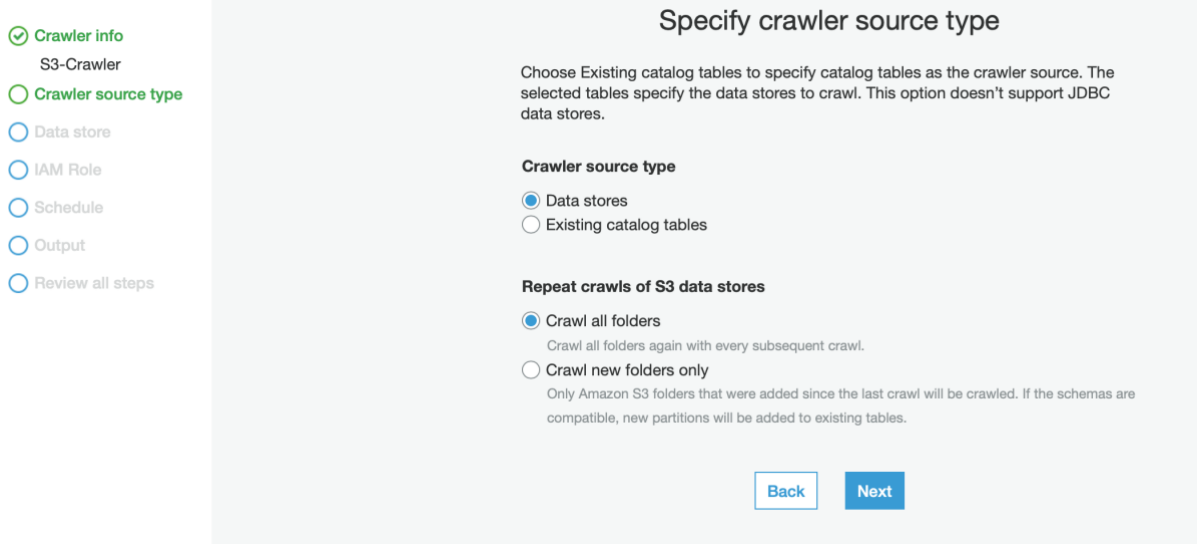
Step 4: Click on Crawlers on the left panel and add a new Crawler



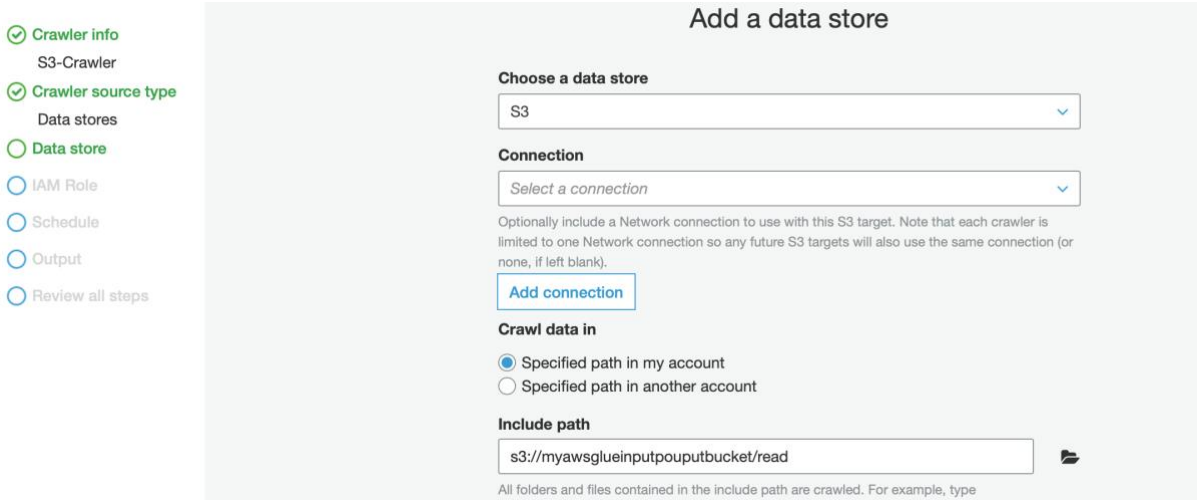
Step 5: Give your crawler a name like S3-Crawler. Click Next



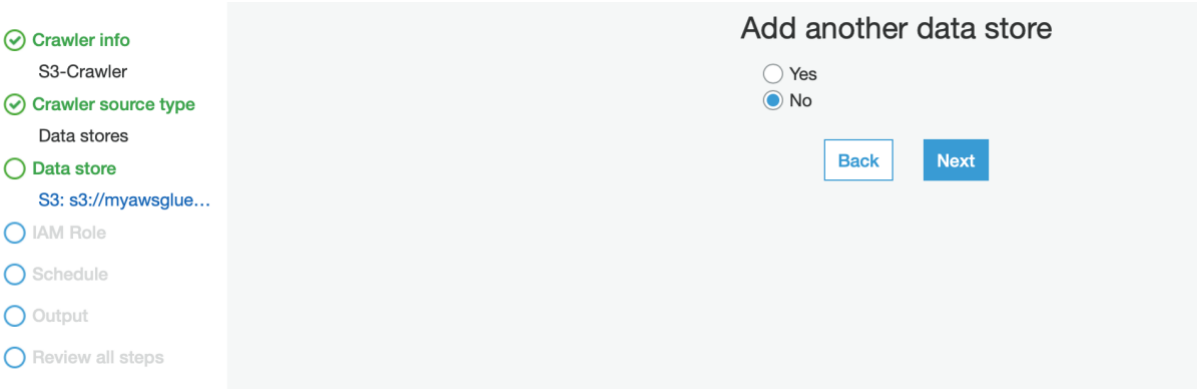
Step 6: Use the default options and click Next



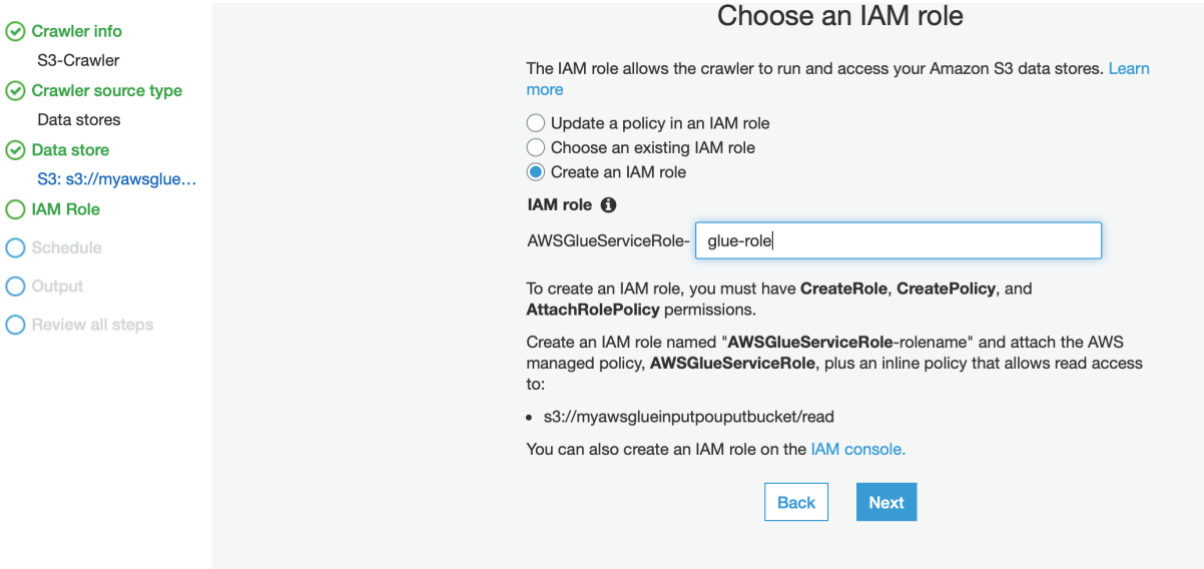
Step 7: Add a data store by selecting the read directory in S3. Click Next



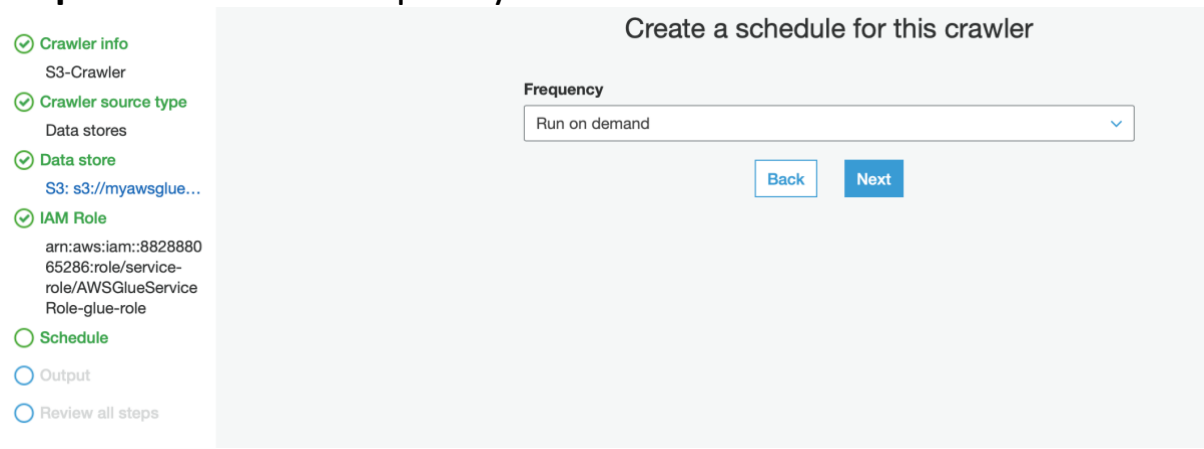
Step 8: Select no in “Add another data store” and click Next



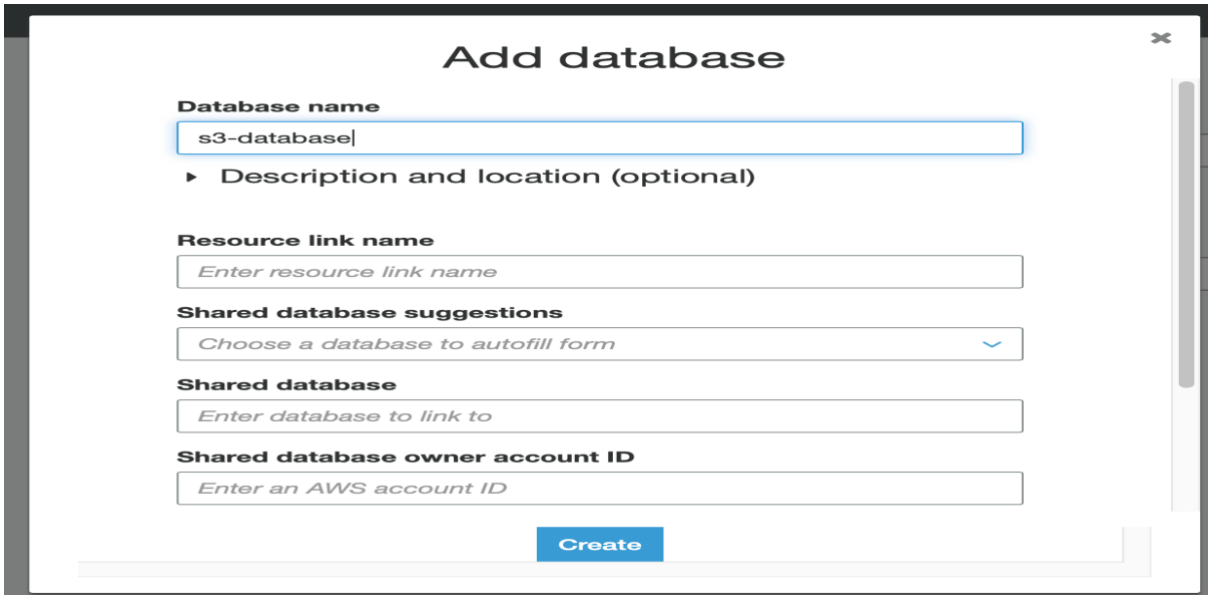
Step 9: Create a new role in IAM and click Next



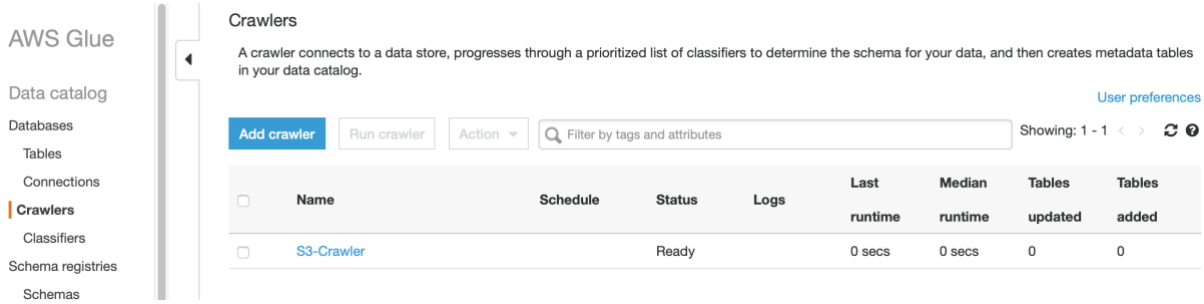
Step 10: Select the frequency as "Run On Demand" and click Next



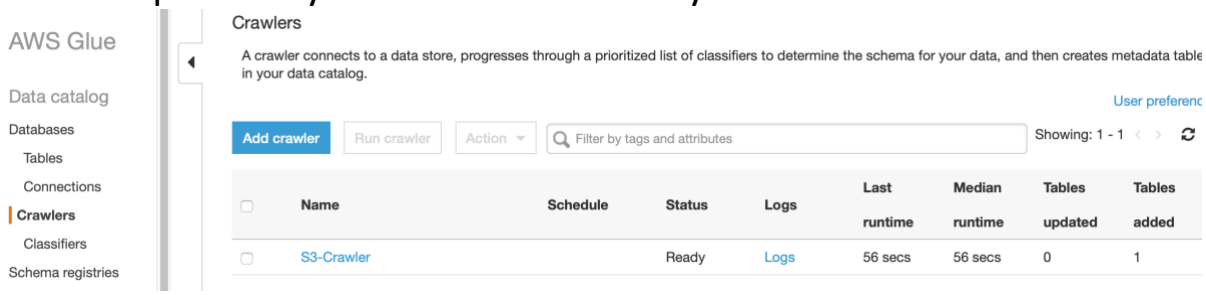
Step 11: Click Add database and enter a database of your choice, click create . Click Next



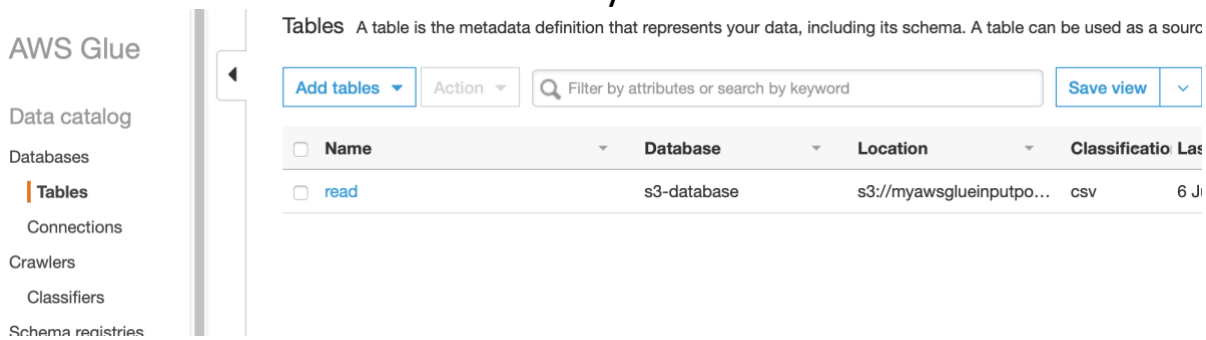
Step 12: Review the steps and click Finish



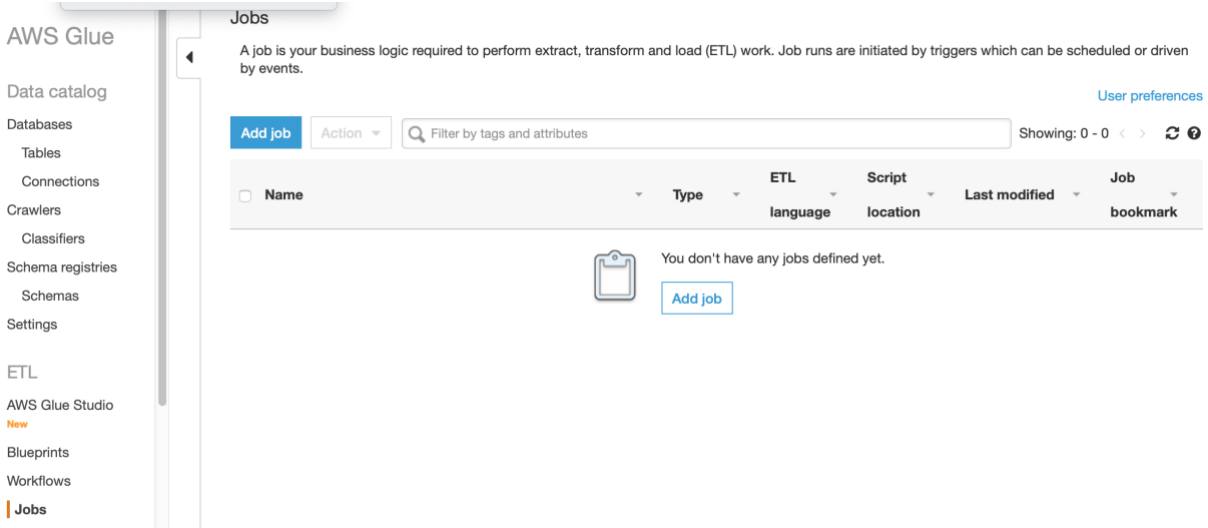
Step 13: Select the crawler created and click Run crawler and wait for it to complete till you see status as ready and Tables added as 1



Step 14: Click database on the left panel and check if the database is created. Also click table to check if your table is created.

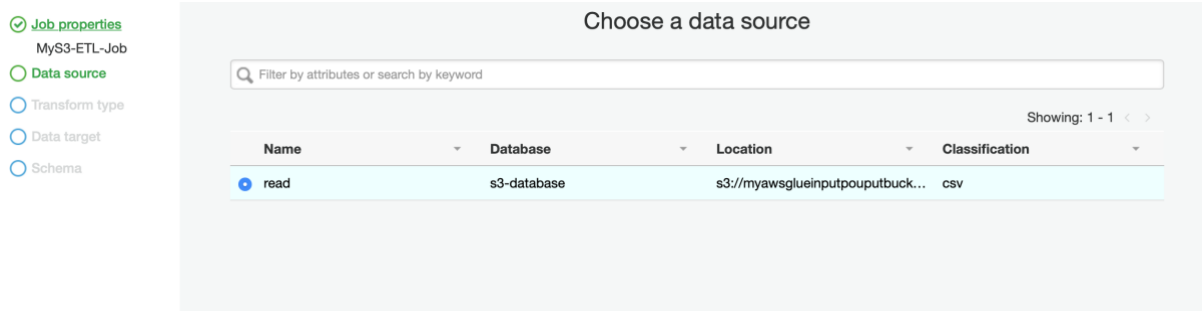


Step 15: Click on the Jobs under the ETL section on the left panel

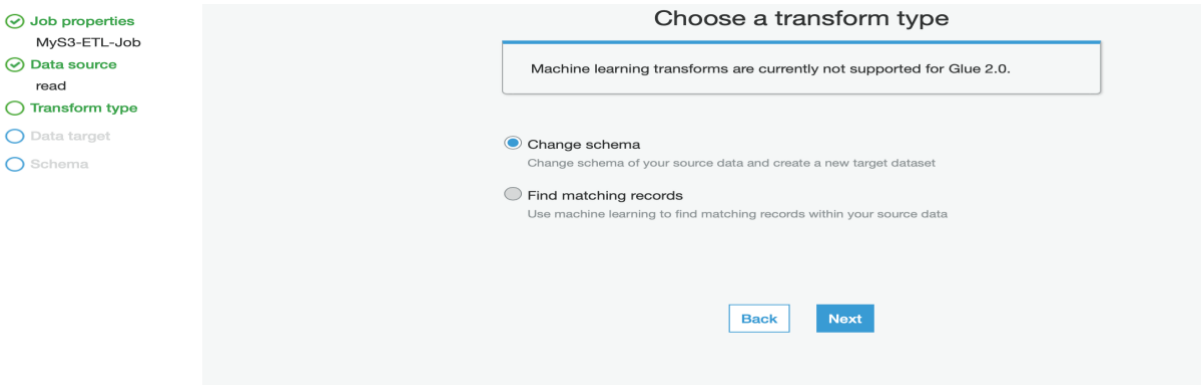


Step 16: Click on Add job and enter the details
Name : MyS3-ETL-Job
IAM Role: Select the available role
Under Security configuration modify
No of workers =2
Job timeout =10
Click Next

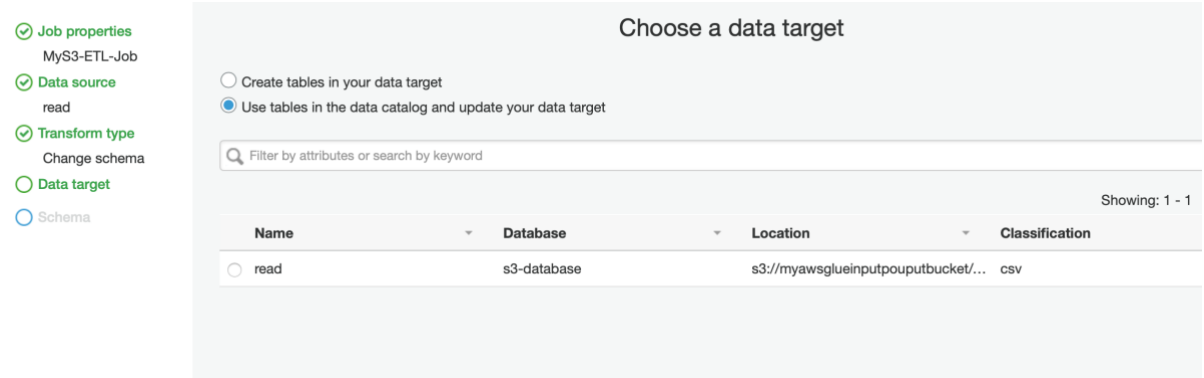
Step 17: Select your data source. Click Next



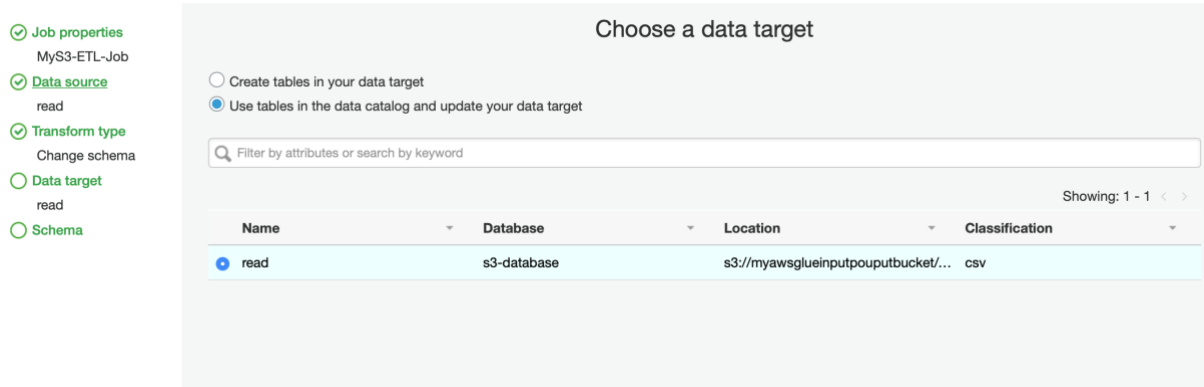
Step 18: Select your Transformation type. Choose default. Click Next



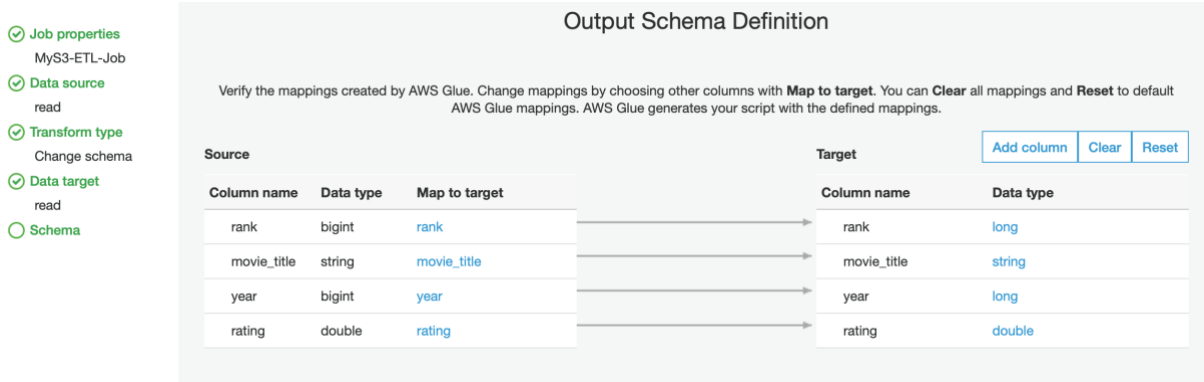
Step 19: Select your data target. Choose the available option .Click Next



Step 20: Select your data target. Choose the available option. Click Next



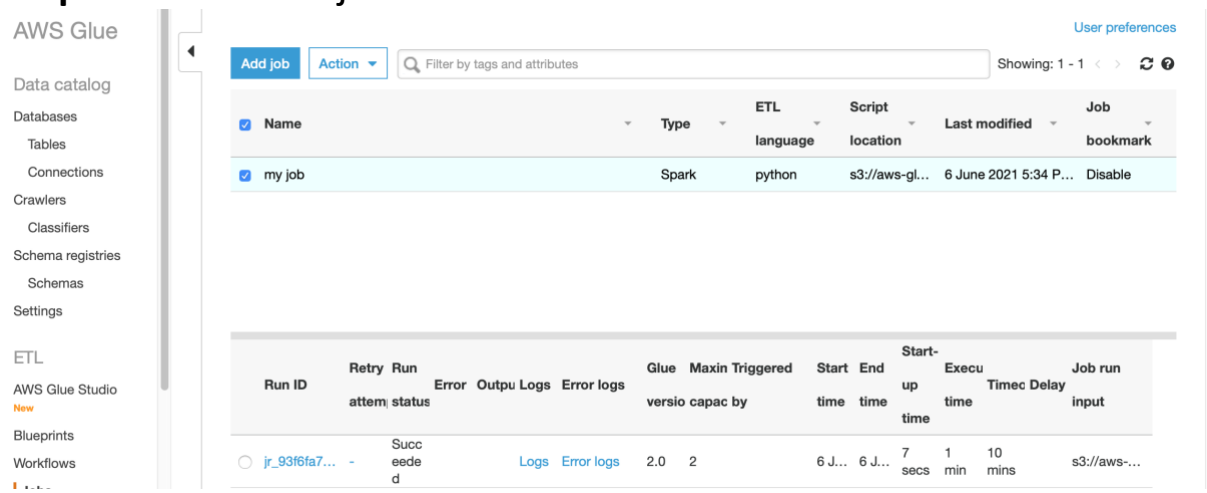
Step 21: Check the output schema definition and click Save job and edit script



Step 22: Replace the script with the one in Code/Glue/glue_transformation.py and Click Save. Gibe the correct database, input folder and output folder names in the script. Click Run Job.

Step 22.1: If you see an error “Job Failed”, the most probable reason would be that Glue does not have access to S3 to write data. So, go to IAM and in roles, search for the role you created in the Step 9 and add a policy S3Full Access and run the Job again.

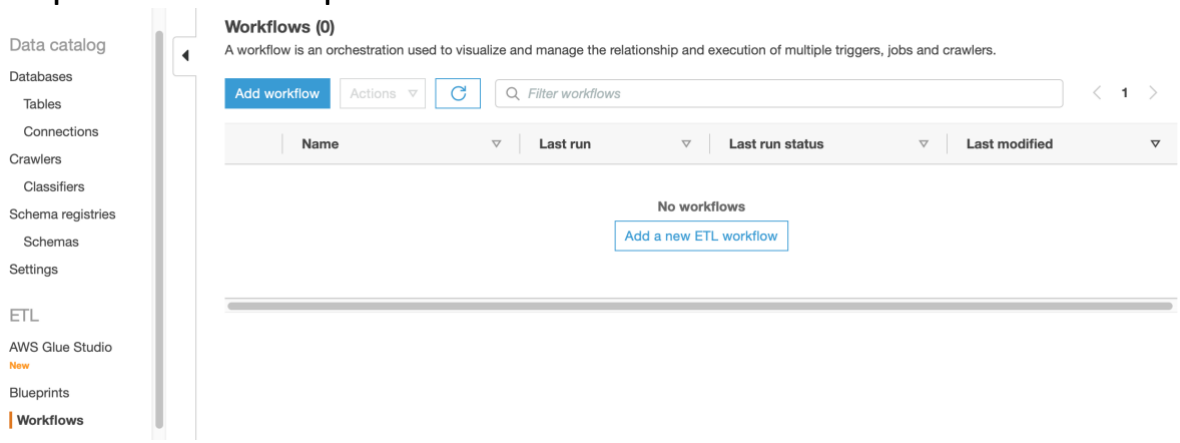
Step 23: Check the job status and make sure it is success.



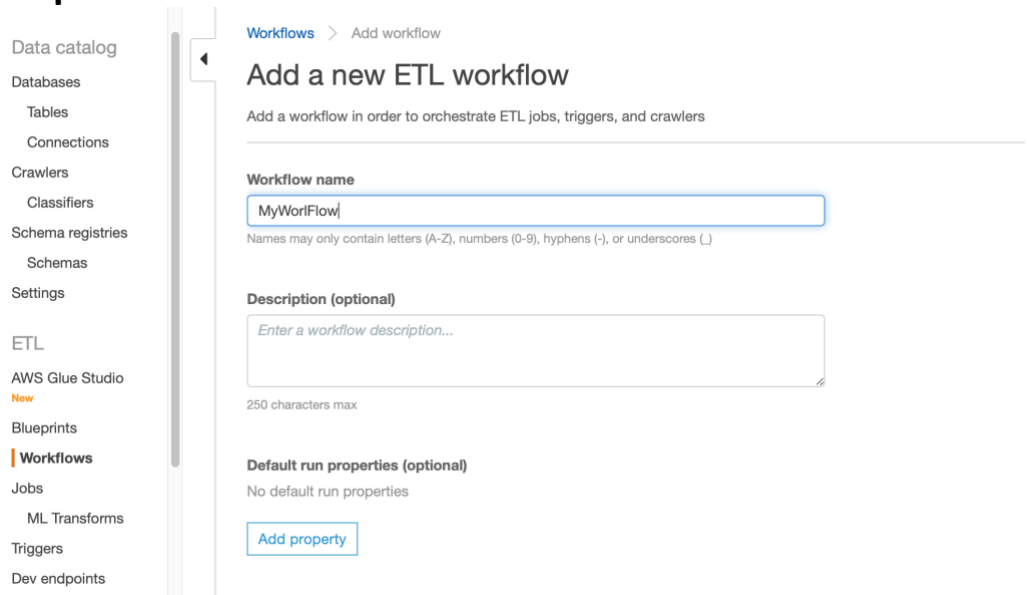
Step 24: Go to your bucket and check the output directory. It should contain a file which has the results

Adding a Workflow

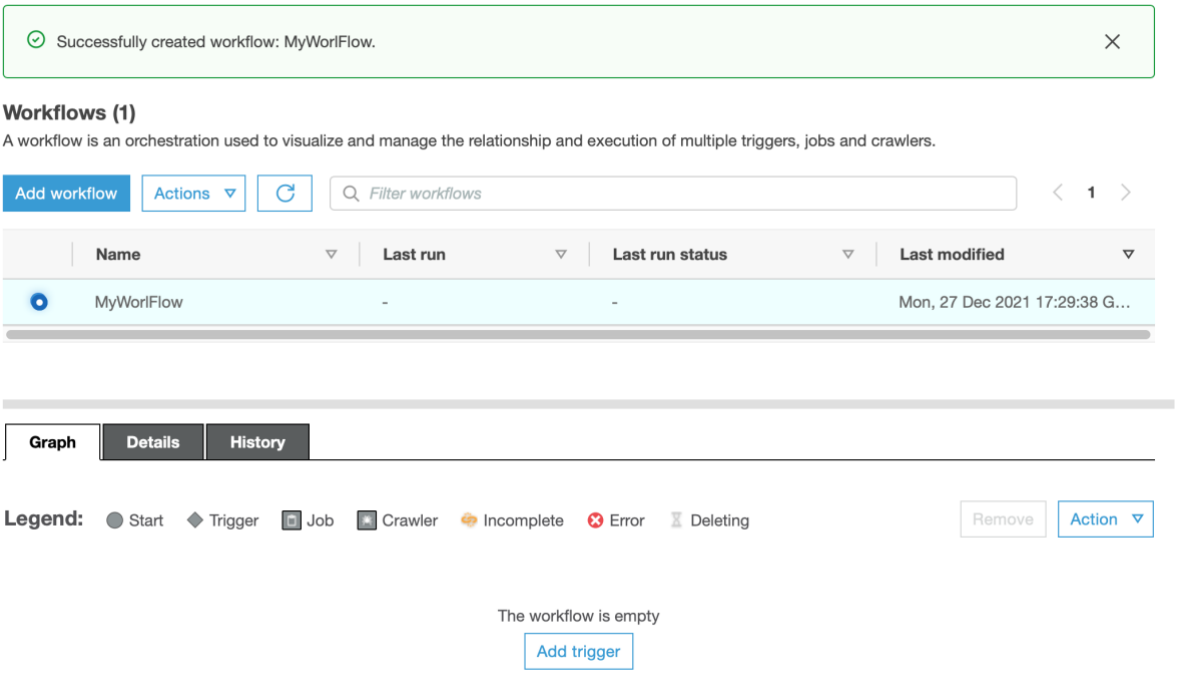
Step 1: On the left panel click Work flow. Click Add workflow



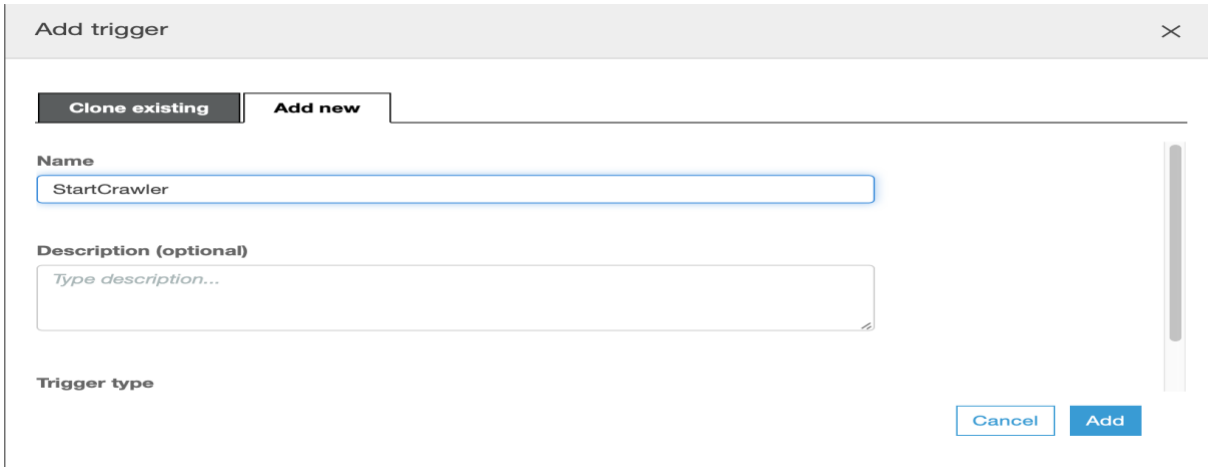
Step 2: Enter work flow name and click Add Workflow



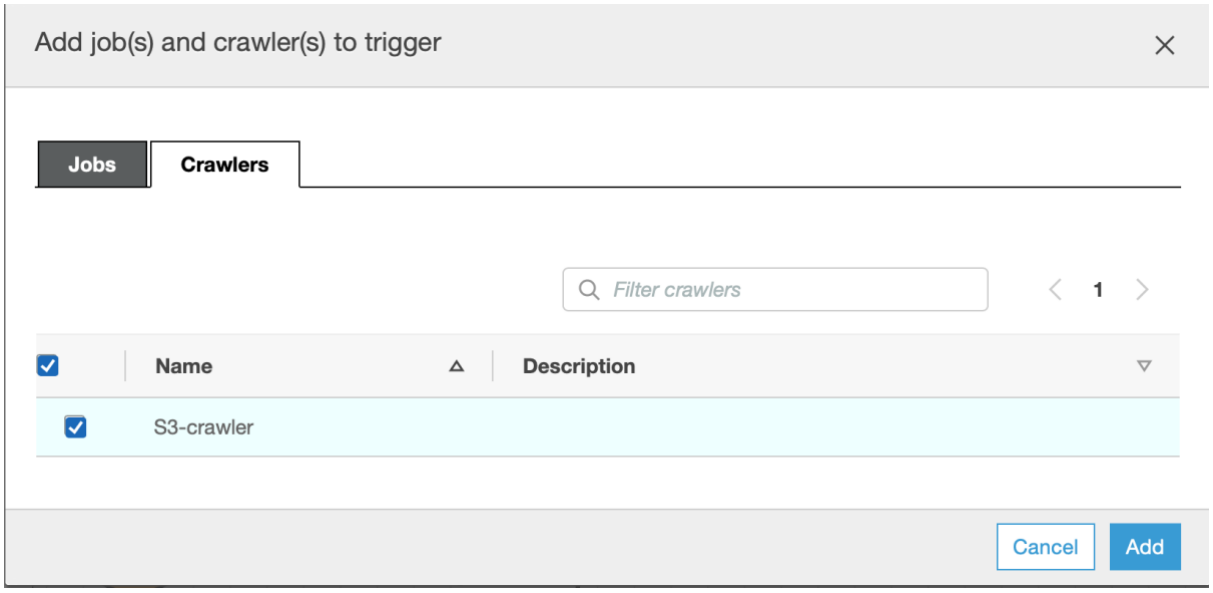
Step 3: Click Add Trigger



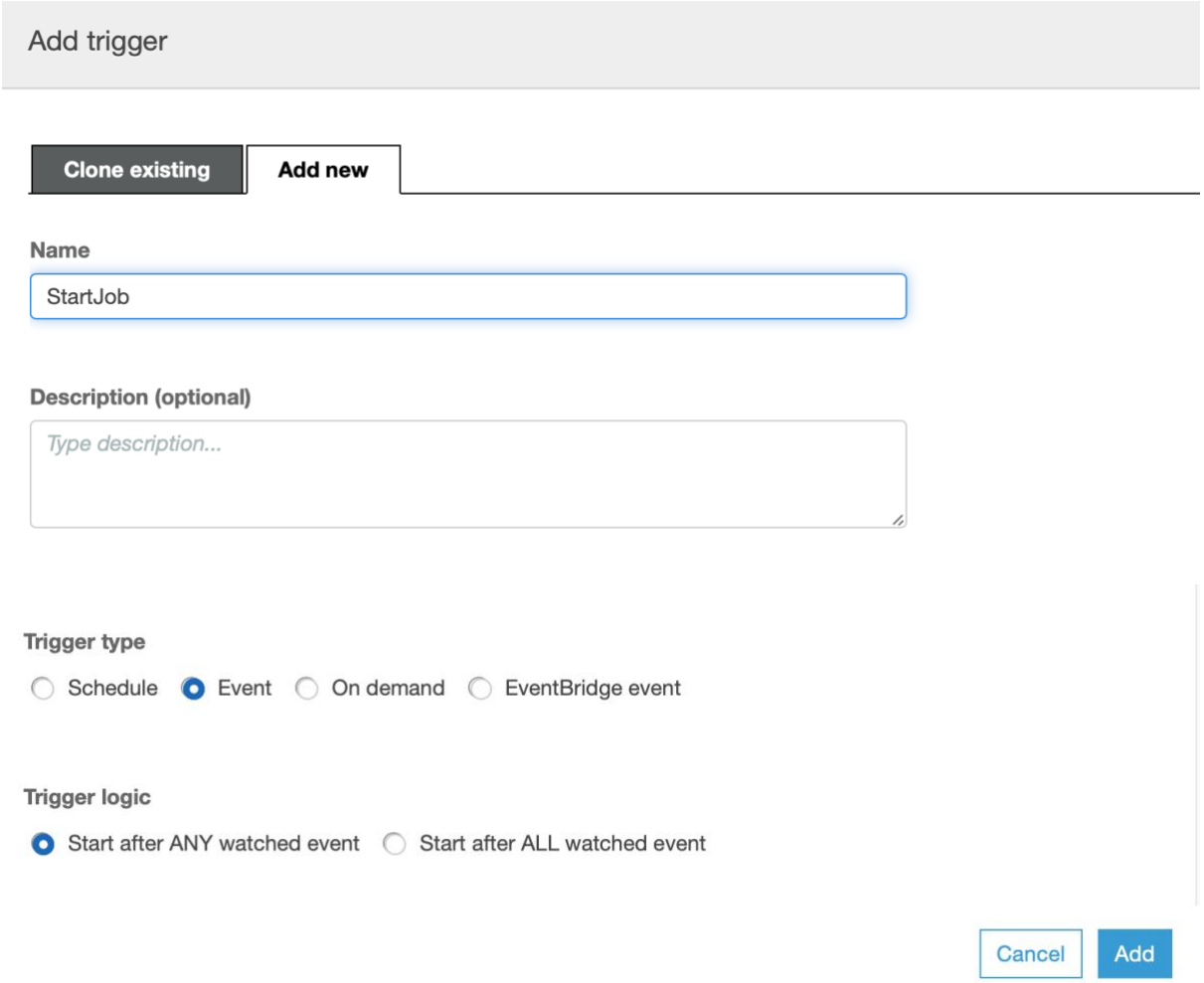
Step 4: Give trigger a name and select OnDemand for trigger type and click Add



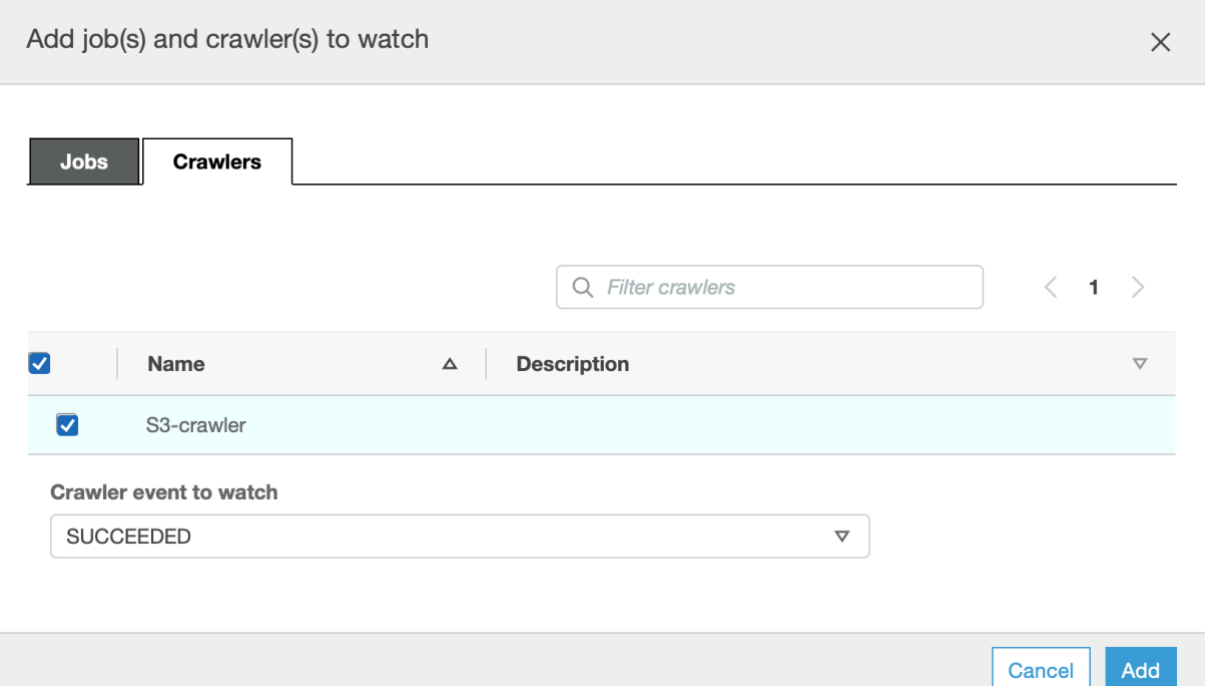
Step 5: Select the Node which pops up a window to add a crawler. Select the crawler tab and select the crawler. Click Add



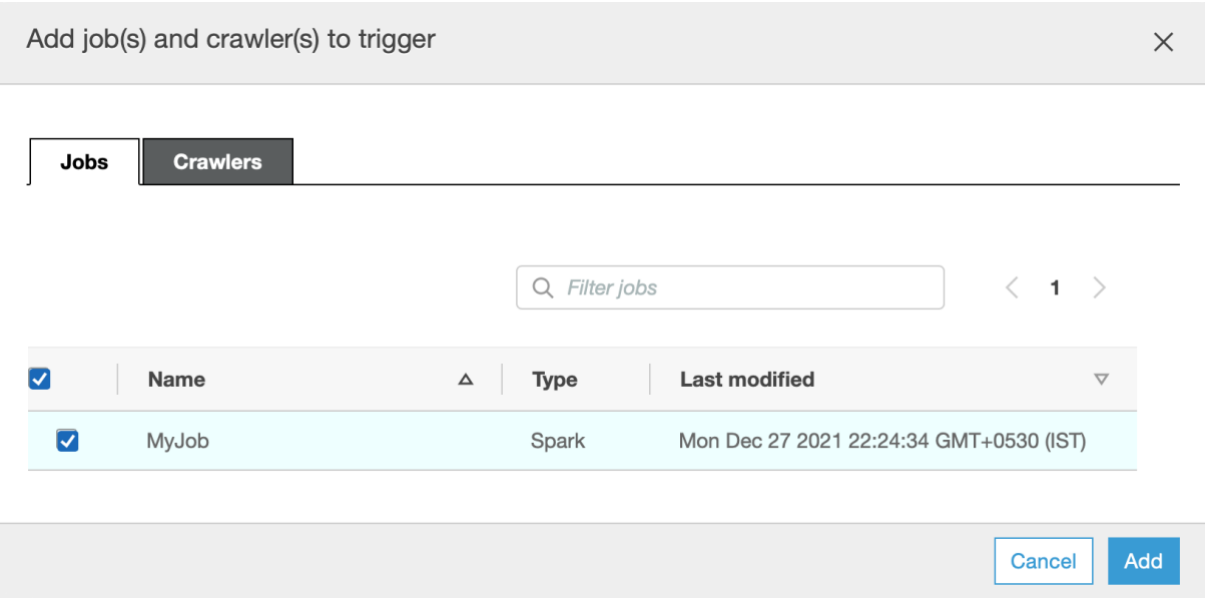
Step 6: Click Actions and select Add Trigger. Give a job name and click Add



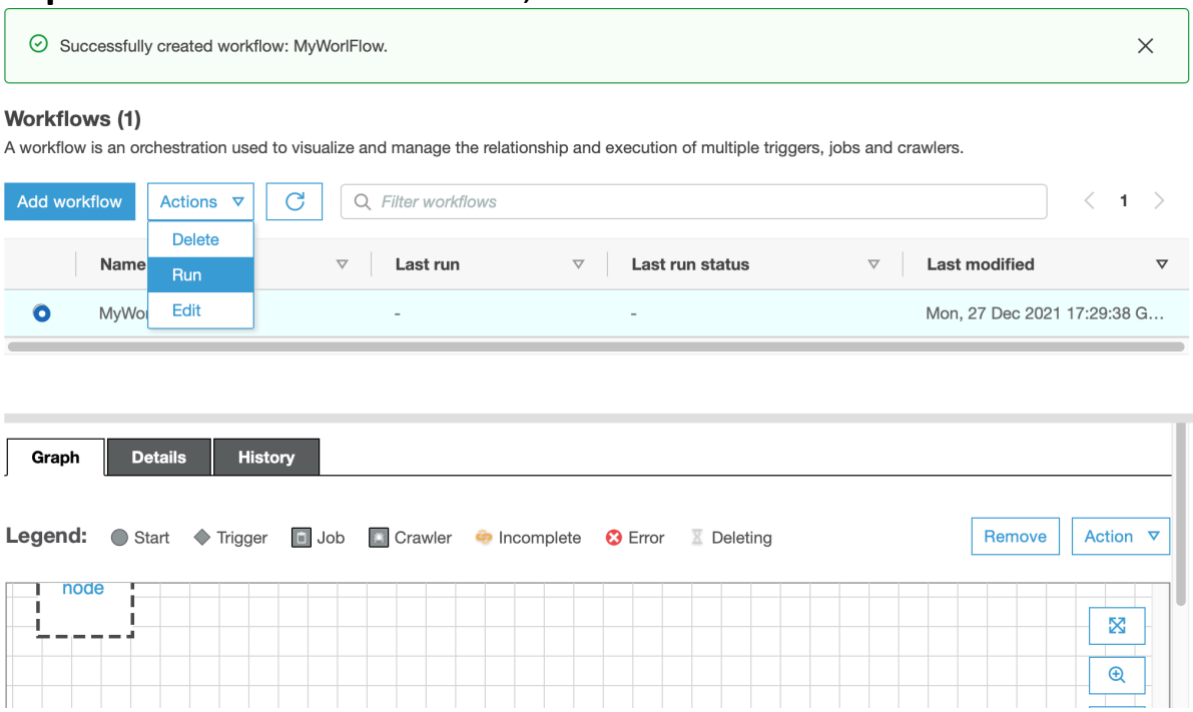
Step 7: Select the Any trigger and click actions and select Add Jobs/crawlers to watch. Select your crawler under the Crawlers and click Add



Step 8: Click on the Node after the Start Job and select your job under the Jobs Tab. Click Add



Step 9: From the actions menu, click on Run



Step 10: Wait till the works goes to completed state and click the output directory in S3